

PATENT APPLICATION

**COMPUTER-AIDED VISUALIZATION OF EXPRESSION
COMPARISON**

*Sub
B1*
Inventor(s): David H. Mack, a citizen of The United States, residing at
2076 Monterey Avenue
Menlo Park, CA 94025

Kurt Gish, a citizen of The United States, residing at
953 Helen Avenue, Apt. 7
Sunnyvale, CA 94086

David Balaban, a citizen of The United States, residing at
7127 Glenview Drive
San Jose, CA 95120

Elina Khurgin, a citizen of The United States, residing at
22999 Voss Avenue
Cupertino, CA 95014

Josie Dai, a citizen of China, residing at
4009 Higuera Road
San Jose, CA 95148

Jim Snyder, a citizen of The United States, residing at
321 Fulton Street
Palo Alto, CA 94301

Assignee: Affymetrix, Inc.
3380 Central Expressway
Santa Clara, CA 95051

Entity: Large
TOWNSEND and TOWNSEND and CREW LLP
Two Embarcadero Center, 8th Floor
San Francisco, California 94111-3834
Tel: 650-326-2400

BACKGROUND OF THE INVENTION

5 The present invention relates to the field of computer systems. More specifically, the present invention relates to computer systems for visualizing analysis results.

10 Devices and computer systems for forming and using arrays of materials on a substrate are known. For example, PCT Publication No. WO 92/10588, incorporated herein by reference for all purposes, describes techniques for sequencing or sequence checking nucleic acids and other materials. Arrays for performing these operations may be formed according to the methods of, for example, the pioneering techniques disclosed in U.S. Patent No. 5,143,854 and U.S. Patent No. 5,593,839 both incorporated herein by reference for all purposes.

15 According to one aspect of the techniques described therein, an array of nucleic acid probes is fabricated at known locations on a substrate or chip. A fluorescently labeled nucleic acid is then brought into contact with the chip and a scanner generates an image file (which is processed into a cell file) indicating the locations where the labeled nucleic acids bound to the chip. Based upon the cell file and identities of the 20 probes at specific locations, it becomes possible to extract information such as the monomer sequence of DNA or RNA. Such systems have been used to form, for example, arrays of DNA that may be used to study and detect mutations relevant to cystic fibrosis, the P53 gene (relevant to certain cancers), HIV, and other genetic characteristics.

25 Computer-aided techniques for monitoring gene expression using such arrays of probes have also been developed as disclosed in U.S. Patent Application No. 08/828,952 (Attorney Docket No. 16528X-028900US) and PCT Publication No. WO 97/10365 (Attorney Docket No. 16528X-017110PC), the contents of which are herein incorporated by reference. Many disease states are characterized by differences in 30 the expression levels of various genes either through changes in the copy number of the genetic DNA or through changes in levels of transcription (e.g., through control of initiation, provision of RNA precursors, RNA processing, etc.) of particular genes. For example, losses and gains of genetic material play an important role in malignant

transformation and progression. Furthermore, changes in the expression (transcription) levels of particular genes (*e.g.*, oncogenes or tumor suppressors), serve as signposts for the presence and progression of various cancers.

It is desirable to identify genes having expression levels relevant to
5 diagnosis of a diseased state by analyzing the expression levels of large numbers of genes in both diseased and normal individuals. Methods for collecting the expression level information have been developed. However, the user interfaces for gene expression monitoring systems that have been developed until now are designed to clearly present the expression of particular pre-selected genes. A user seeking to identify, *e.g.*, an oncogene
10 or a tumor suppressor gene, must individually review the expression level of large numbers of genes and compare the expression levels between diseased and normal individuals. What is needed is a user interface that takes advantage of collected gene expression information to help the user to identify particular genes of interest.
15
20

SUMMARY OF THE INVENTION

The present invention provides innovative systems and methods for visualizing information collected from analyzing samples. The samples may include nucleic acids, proteins, or other polymers. Gene expression level as determined from analysis of a nucleic acid sample is one possible analysis result that may be visualized. In one embodiment, a computer system may display the expression levels of multiple genes simultaneously in a way that facilitates user identification of genes whose expression is significant to a characteristic such as disease or resistance to disease. Additionally, the computer system may facilitate display of further information about relevant genes once they are identified.
25

A first aspect of the invention provides a computer-implemented method for presenting expression level information as collected from first and second samples. The method includes steps of: displaying a first axis corresponding to expression level in the first sample, and displaying a second axis substantially perpendicular to the first axis, the second axis corresponding to expression level in the second sample. The method further
30 includes a step of: for a selected expressed sequence, displaying a mark at a position. The position is selected relative to the first axis in accordance with an expression level of the selected expressed sequence in the first sample and relative to the second axis in accordance with an expression level of the selected expressed sequence in the second

sample. A particularly useful application is displaying many marks simultaneously for many selected genes to discover which ones of the selected genes may be relevant to the characteristic.

A second aspect of the invention provides a computer-implemented method
5 of presenting sample analysis information. The method includes steps of: displaying a first axis corresponding to a concentration of a compound in a first sample as determined by monitoring binding of the compound to a selected polymer having binding affinity to the compound, and displaying a second axis substantially perpendicular to the first axis. The second axis corresponds to a concentration of the compound in the second sample as
10 determined by monitoring binding of the compound to the selected polymer. The method further preferably includes a step of displaying a mark at a position. The position is selected relative to the first axis in accordance with the concentration in the first sample and relative to the second axis in accordance with the concentration in the second sample.

A further understanding of the nature and advantages of the inventions herein may be realized by reference to the remaining portions of the specification and the attached drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 illustrates an example of a computer system that may be used to
20 execute software embodiments of the present invention.

Fig. 2 shows a system block diagram of a typical computer system.

Fig. 3 illustrates an overall system for forming and analyzing arrays of polymers including biological materials such as DNA or RNA.

Fig. 4 is an illustration of an embodiment of software for the overall
25 system.

Fig. 5 shows a flowchart of a process of monitoring the expression of a gene by comparing hybridization intensities of pairs of perfect match and mismatch probes.

Fig. 6 shows a screen display illustrating gene expression levels for
30 multiple genes as collected from both normal and diseased tissue.

Sulf
Figs. 7A-7B show screen displays illustrating information about a particular gene selected from the display of Fig. 6.

DESCRIPTION OF SPECIFIC EMBODIMENTS

The present invention provides innovative methods of monitoring visualizing gene expression. In the description that follows, the invention will be described in reference to preferred embodiments. However, the description is provided for purposes of illustration and not for limiting the spirit and scope of the invention.

Fig. 1 illustrates an example of a computer system that may be used to execute software embodiments of the present invention. Fig. 1 shows a computer system 1 which includes a monitor 3, screen 5, cabinet 7, keyboard 9, and mouse 11. Mouse 11 may have one or more buttons such as mouse buttons 13. Cabinet 7 houses a CD-ROM drive 15 and a hard drive (not shown) that may be utilized to store and retrieve software programs including computer code incorporating the present invention. Although a CD-ROM 17 is shown as the computer readable medium, other computer readable media including floppy disks, DRAM, hard drives, flash memory, tape, and the like may be utilized. Cabinet 7 also houses familiar computer components (not shown) such as a processor, memory, and the like.

Fig. 2 shows a system block diagram of computer system 1 used to execute software embodiments of the present invention. As in Fig. 1, computer system 1 includes monitor 3 and keyboard 9. Computer system 1 further includes subsystems such as a central processor 50, system memory 52, I/O controller 54, display adapter 56, removable disk 58, fixed disk 60, network interface 62, and speaker 64. Removable disk 58 is representative of removable computer readable media like floppies, tape, CD-ROM, removable hard drive, flash memory, and the like. Fixed disk 60 is representative of an internal hard drive or the like. Other computer systems suitable for use with the present invention may include additional or fewer subsystems. For example, another computer system could include more than one processor 50 (*i.e.*, a multi-processor system) or memory cache.

Arrows such as 66 represent the system bus architecture of computer system 1. However, these arrows are illustrative of any interconnection scheme serving to link the subsystems. For example, display adapter 56 may be connected to central processor 50 through a local bus or the system may include a memory cache. Computer system 1 shown in Fig. 2 is but an example of a computer system suitable for use with the present invention. Other configurations of subsystems suitable for use with the present invention will be readily apparent to one of ordinary skill in the art. In one

embodiment, the computer system is an IBM compatible personal computer.

The VLSIPS™ and GeneChip™ technologies provide methods of making and using very large arrays of polymers, such as nucleic acids, on very small chips. See U.S. Patent No. 5,143,854 and PCT Patent Publication Nos. WO 90/15070 and 92/10092,

- 5 each of which is hereby incorporated by reference for all purposes. Nucleic acid probes on the chip are used to detect complementary nucleic acid sequences in a sample nucleic acid of interest (the "target" nucleic acid).

It should be understood that the probes need not be nucleic acid probes but may also be other receptors, such as antibodies, or polymers such as peptides. Peptide 10 probes may be used to detect the concentration of other peptides, proteins, or other compounds in a sample. The probes must be carefully selected to have bonding affinity to the compound whose concentration they are to be used to measure.

In one embodiment, the present invention provides methods of visualizing information relating to the concentration of compounds in a sample as measured by monitoring affinity of the compounds to probes. In a particular application, the concentration information is generated by analysis of hybridization intensity files for a chip containing hybridized nucleic acid probes. The hybridization of a nucleic acid sample to certain probes may represent the expression level of one more genes or expressed sequence tags (ESTs). The expression level of a gene or EST is herein 15 understood to be the concentration within a sample of mRNA or protein that would result from the transcription of the gene or EST.

Expression level information visualized by virtue of the present invention need not be obtained from probes but may originate from any source. If the expression information is collected from a probe array, the probe array need not meet any particular 20 criteria for size and density. Furthermore, the present invention is not limited to visualizing fluorescent measurements of bondings such as hybridizations but may be readily utilized to visualize other measurements.

Concentration of compounds other than nucleic acids may be visualized according to one embodiment of the present invention. For example, a probe array may 25 include peptide probes which may be exposed to protein samples, polypeptide samples, or other compounds which may or may not bond to the peptide probes. By appropriate selection of the peptide probes, one may detect the presence or absence of particular compounds which would bond to the peptide probes.

For purposes of illustration, the present invention is described as being part of a system that designs a chip mask, synthesizes the probes on the chip, labels nucleic acids from a target sample, and scans the hybridized probes. Such a system is set forth in U.S. Patent No. 5,571,639 which is hereby incorporated by reference for all purposes.

- 5 However, the present invention may be used separately from the overall system for analyzing data generated by such systems, such as at remote locations, or for visualizing the results of other systems for generating expression information, or for visualizing concentrations of polymers other than nucleic acids.

Fig. 3 illustrates a computerized system for forming and analyzing arrays of biological materials such as RNA or DNA. A computer 100 is used to design arrays of biological polymers such as RNA or DNA. The computer 100 may be, for example, an appropriately programmed IBM personal computer compatible running Windows NT including appropriate memory and a CPU as shown in Figs. 1 and 2. The computer system 100 obtains inputs from a user regarding characteristics of a gene of interest, and other inputs regarding the desired features of the array. Optionally, the computer system may obtain information regarding a specific genetic sequence of interest from an external or internal database 102 such as GenBank. The output of the computer system 100 is a set of chip design computer files 104 in the form of, for example, a switch matrix, as described in PCT application WO 92/10092, and other associated computer files.

20 The chip design files are provided to a system 106 that designs the lithographic masks used in the fabrication of arrays of molecules such as DNA. The system or process 106 may include the hardware necessary to manufacture masks 110 and also the necessary computer hardware and software 108 necessary to lay the mask patterns out on the mask in an efficient manner. As with the other features in Fig. 3, 25 such equipment may or may not be located at the same physical site, but is shown together for ease of illustration in Fig. 3. The system 106 generates masks 110 or other synthesis patterns such as chrome-on-glass masks for use in the fabrication of polymer arrays.

The masks 110, as well as selected information relating to the design of the 30 chips from system 100, are used in a synthesis system 112. Synthesis system 112 includes the necessary hardware and software used to fabricate arrays of polymers on a substrate or chip 114. For example, synthesizer 112 includes a light source 116 and a chemical flow cell 118 on which the substrate or chip 114 is placed. Mask 110 is placed

between the light source and the substrate/chip, and the two are translated relative to each other at appropriate times for deprotection of selected regions of the chip. Selected chemical reagents are directed through flow cell 118 for coupling to deprotected regions, as well as for washing and other operations. All operations are preferably directed by an appropriately programmed computer 119, which may or may not be the same computer as the computer(s) used in mask design and mask making.

The substrates fabricated by synthesis system 112 are optionally diced into smaller chips and exposed to marked targets. The targets may or may not be complementary to one or more of the molecules on the substrate. The targets are marked with a label such as a fluorescein label (indicated by an asterisk in Fig. 3) and placed in scanning system 120. Scanning system 120 again operates under the direction of an appropriately programmed digital computer 122, which also may or may not be the same computer as the computers used in synthesis, mask making, and mask design. The scanner 120 includes a detection device 124 such as a confocal microscope or CCD (charge-coupled device) that is used to detect the location where labeled target has bound to the substrate. The output of scanner 120 is an image file(s) 124 indicating, in the case of fluorescein labeled target, the fluorescence intensity (photon counts or other related measurements, such as voltage) as a function of position on the substrate. Since higher photon counts will be observed where the labeled target has bound more strongly to the array of polymers, and since the monomer sequence of the polymers on the substrate is known as a function of position, it becomes possible to determine the sequence(s) of polymer(s) on the substrate that are complementary to the target.

The image file 124 is provided as input to an analysis system 126 that incorporates the visualization and analysis methods of the present invention. Again, the analysis system may be any one of a wide variety of computer system. The present invention provides various methods of analyzing and visualizing the chip design files and the image files, providing appropriate output 128. The chip design need not include any particular number of probes. It should be understood that the present invention does not require any particular source of expression level information.

Fig. 4 provides a simplified illustration of the overall software system used in the operation of one embodiment of the invention. As shown in Fig. 4, the system first identifies the nucleotide sequence(s) or targets that would be of interest in a particular expression level analysis at step 202. The sequences of interest correspond to

mRNA transcripts of one or more genes, ESTs or nucleic acids derived from the mRNA transcripts. Sequence selection may be provided via manual input of text files or may be from external sources such as GenBank.

At step 204 the system evaluates the sequences of interest to determine or assist the user in determining which probes would be desirable on the chip, and provides an appropriate "layout" on the chip for the probes. The process of selecting probes for an expression level analysis is explained in PCT Publication No. WO 97/10365, the contents of which are herein incorporated by reference. An alternative probe selection process that does not require prior knowledge of sequences of interest is explained in PCT Publication No. WO97/27317 (Attorney Docket No. 18547-019410PC), the contents of which are herein incorporated by reference. Further general background on probe selection is found in PCT Publication No. WO95/11995 (Attorney Docket No. 18547-004111PC) and PCT Publication No. WO97/29212 (Attorney Docket No. 18547-018540PC), the contents of which are herein incorporated by reference. The term "perfect match probe" refers to a probe that has a sequence that is perfectly complementary to a particular target sequence. The test probe is typically perfectly complementary to a portion (subsequence) of the target sequence. The term "mismatch control" or "mismatch probe" refer to probes whose sequence is deliberately selected not to be perfectly complementary to a particular target sequence. For each mismatch (MM) control in an array there typically exists a corresponding perfect match (PM) probe that is perfectly complementary to the same particular target sequence.

The process compares hybridization intensities of pairs of perfect match and mismatch probes that are preferably covalently attached to the surface of a substrate or chip. Most preferably, the nucleic acid probes have a density greater than about 60 different nucleic acid probes per 1 cm² of the substrate.

Initially, nucleic acid probes are selected that are complementary to the target sequence. These probes are the perfect match probes. Another set of probes is specified that are intended to be not perfectly complementary to the target sequence. These probes are the mismatch probes and each mismatch probe includes at least one nucleotide mismatch from a perfect match probe. Accordingly, a mismatch probe and the perfect match probe to which it is identical except for one base make up a pair. As mentioned earlier, the nucleotide mismatch is preferably near the center of the mismatch probe.

The probe lengths of the perfect match probes are typically chosen to exhibit detectably greater hybridization with the target sequence relative to the mismatch probes. For example, the nucleic acid probes may be all 20-mers. However, probes of varying lengths may also be synthesized on the substrate for any number of reasons
5 including resolving ambiguities.

Again referring to Fig. 4, at step 206 the masks for the synthesis are designed. At step 208 the software utilizes the mask design and layout information to make the DNA or other polymer chips. This step 208 will control, among other things, relative translation of a substrate and the mask, the flow of desired reagents through a flow cell, the synthesis temperature of the flow cell, and other parameters. At step 210, another piece of software is used in scanning a chip thus synthesized and exposed to a labeled target. The software controls the scanning of the chip, and stores the data thus obtained in a file that may later be utilized to extract hybridization information.
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
110
115
120
125
130
135
140
145
150
155
160
165
170
175
180
185
190
195
200
205
210
215
220
225
230
235
240
245
250
255
260
265
270
275
280
285
290
295
300
305
310
315
320
325
330
335
340
345
350
355
360
365
370
375
380
385
390
395
400
405
410
415
420
425
430
435
440
445
450
455
460
465
470
475
480
485
490
495
500
505
510
515
520
525
530
535
540
545
550
555
560
565
570
575
580
585
590
595
600
605
610
615
620
625
630
635
640
645
650
655
660
665
670
675
680
685
690
695
700
705
710
715
720
725
730
735
740
745
750
755
760
765
770
775
780
785
790
795
800
805
810
815
820
825
830
835
840
845
850
855
860
865
870
875
880
885
890
895
900
905
910
915
920
925
930
935
940
945
950
955
960
965
970
975
980
985
990
995
1000
1005
1010
1015
1020
1025
1030
1035
1040
1045
1050
1055
1060
1065
1070
1075
1080
1085
1090
1095
1100
1105
1110
1115
1120
1125
1130
1135
1140
1145
1150
1155
1160
1165
1170
1175
1180
1185
1190
1195
1200
1205
1210
1215
1220
1225
1230
1235
1240
1245
1250
1255
1260
1265
1270
1275
1280
1285
1290
1295
1300
1305
1310
1315
1320
1325
1330
1335
1340
1345
1350
1355
1360
1365
1370
1375
1380
1385
1390
1395
1400
1405
1410
1415
1420
1425
1430
1435
1440
1445
1450
1455
1460
1465
1470
1475
1480
1485
1490
1495
1500
1505
1510
1515
1520
1525
1530
1535
1540
1545
1550
1555
1560
1565
1570
1575
1580
1585
1590
1595
1600
1605
1610
1615
1620
1625
1630
1635
1640
1645
1650
1655
1660
1665
1670
1675
1680
1685
1690
1695
1700
1705
1710
1715
1720
1725
1730
1735
1740
1745
1750
1755
1760
1765
1770
1775
1780
1785
1790
1795
1800
1805
1810
1815
1820
1825
1830
1835
1840
1845
1850
1855
1860
1865
1870
1875
1880
1885
1890
1895
1900
1905
1910
1915
1920
1925
1930
1935
1940
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
2025
2030
2035
2040
2045
2050
2055
2060
2065
2070
2075
2080
2085
2090
2095
2100
2105
2110
2115
2120
2125
2130
2135
2140
2145
2150
2155
2160
2165
2170
2175
2180
2185
2190
2195
2200
2205
2210
2215
2220
2225
2230
2235
2240
2245
2250
2255
2260
2265
2270
2275
2280
2285
2290
2295
2300
2305
2310
2315
2320
2325
2330
2335
2340
2345
2350
2355
2360
2365
2370
2375
2380
2385
2390
2395
2400
2405
2410
2415
2420
2425
2430
2435
2440
2445
2450
2455
2460
2465
2470
2475
2480
2485
2490
2495
2500
2505
2510
2515
2520
2525
2530
2535
2540
2545
2550
2555
2560
2565
2570
2575
2580
2585
2590
2595
2600
2605
2610
2615
2620
2625
2630
2635
2640
2645
2650
2655
2660
2665
2670
2675
2680
2685
2690
2695
2700
2705
2710
2715
2720
2725
2730
2735
2740
2745
2750
2755
2760
2765
2770
2775
2780
2785
2790
2795
2800
2805
2810
2815
2820
2825
2830
2835
2840
2845
2850
2855
2860
2865
2870
2875
2880
2885
2890
2895
2900
2905
2910
2915
2920
2925
2930
2935
2940
2945
2950
2955
2960
2965
2970
2975
2980
2985
2990
2995
3000
3005
3010
3015
3020
3025
3030
3035
3040
3045
3050
3055
3060
3065
3070
3075
3080
3085
3090
3095
3100
3105
3110
3115
3120
3125
3130
3135
3140
3145
3150
3155
3160
3165
3170
3175
3180
3185
3190
3195
3200
3205
3210
3215
3220
3225
3230
3235
3240
3245
3250
3255
3260
3265
3270
3275
3280
3285
3290
3295
3300
3305
3310
3315
3320
3325
3330
3335
3340
3345
3350
3355
3360
3365
3370
3375
3380
3385
3390
3395
3400
3405
3410
3415
3420
3425
3430
3435
3440
3445
3450
3455
3460
3465
3470
3475
3480
3485
3490
3495
3500
3505
3510
3515
3520
3525
3530
3535
3540
3545
3550
3555
3560
3565
3570
3575
3580
3585
3590
3595
3600
3605
3610
3615
3620
3625
3630
3635
3640
3645
3650
3655
3660
3665
3670
3675
3680
3685
3690
3695
3700
3705
3710
3715
3720
3725
3730
3735
3740
3745
3750
3755
3760
3765
3770
3775
3780
3785
3790
3795
3800
3805
3810
3815
3820
3825
3830
3835
3840
3845
3850
3855
3860
3865
3870
3875
3880
3885
3890
3895
3900
3905
3910
3915
3920
3925
3930
3935
3940
3945
3950
3955
3960
3965
3970
3975
3980
3985
3990
3995
4000
4005
4010
4015
4020
4025
4030
4035
4040
4045
4050
4055
4060
4065
4070
4075
4080
4085
4090
4095
4100
4105
4110
4115
4120
4125
4130
4135
4140
4145
4150
4155
4160
4165
4170
4175
4180
4185
4190
4195
4200
4205
4210
4215
4220
4225
4230
4235
4240
4245
4250
4255
4260
4265
4270
4275
4280
4285
4290
4295
4300
4305
4310
4315
4320
4325
4330
4335
4340
4345
4350
4355
4360
4365
4370
4375
4380
4385
4390
4395
4400
4405
4410
4415
4420
4425
4430
4435
4440
4445
4450
4455
4460
4465
4470
4475
4480
4485
4490
4495
4500
4505
4510
4515
4520
4525
4530
4535
4540
4545
4550
4555
4560
4565
4570
4575
4580
4585
4590
4595
4600
4605
4610
4615
4620
4625
4630
4635
4640
4645
4650
4655
4660
4665
4670
4675
4680
4685
4690
4695
4700
4705
4710
4715
4720
4725
4730
4735
4740
4745
4750
4755
4760
4765
4770
4775
4780
4785
4790
4795
4800
4805
4810
4815
4820
4825
4830
4835
4840
4845
4850
4855
4860
4865
4870
4875
4880
4885
4890
4895
4900
4905
4910
4915
4920
4925
4930
4935
4940
4945
4950
4955
4960
4965
4970
4975
4980
4985
4990
4995
5000
5005
5010
5015
5020
5025
5030
5035
5040
5045
5050
5055
5060
5065
5070
5075
5080
5085
5090
5095
5100
5105
5110
5115
5120
5125
5130
5135
5140
5145
5150
5155
5160
5165
5170
5175
5180
5185
5190
5195
5200
5205
5210
5215
5220
5225
5230
5235
5240
5245
5250
5255
5260
5265
5270
5275
5280
5285
5290
5295
5300
5305
5310
5315
5320
5325
5330
5335
5340
5345
5350
5355
5360
5365
5370
5375
5380
5385
5390
5395
5400
5405
5410
5415
5420
5425
5430
5435
5440
5445
5450
5455
5460
5465
5470
5475
5480
5485
5490
5495
5500
5505
5510
5515
5520
5525
5530
5535
5540
5545
5550
5555
5560
5565
5570
5575
5580
5585
5590
5595
5600
5605
5610
5615
5620
5625
5630
5635
5640
5645
5650
5655
5660
5665
5670
5675
5680
5685
5690
5695
5700
5705
5710
5715
5720
5725
5730
5735
5740
5745
5750
5755
5760
5765
5770
5775
5780
5785
5790
5795
5800
5805
5810
5815
5820
5825
5830
5835
5840
5845
5850
5855
5860
5865
5870
5875
5880
5885
5890
5895
5900
5905
5910
5915
5920
5925
5930
5935
5940
5945
5950
5955
5960
5965
5970
5975
5980
5985
5990
5995
6000
6005
6010
6015
6020
6025
6030
6035
6040
6045
6050
6055
6060
6065
6070
6075
6080
6085
6090
6095
6100
6105
6110
6115
6120
6125
6130
6135
6140
6145
6150
6155
6160
6165
6170
6175
6180
6185
6190
6195
6200
6205
6210
6215
6220
6225
6230
6235
6240
6245
6250
6255
6260
6265
6270
6275
6280
6285
6290
6295
6300
6305
6310
6315
6320
6325
6330
6335
6340
6345
6350
6355
6360
6365
6370
6375
6380
6385
6390
6395
6400
6405
6410
6415
6420
6425
6430
6435
6440
6445
6450
6455
6460
6465
6470
6475
6480
6485
6490
6495
6500
6505
6510
6515
6520
6525
6530
6535
6540
6545
6550
6555
6560
6565
6570
6575
6580
6585
6590
6595
6600
6605
6610
6615
6620
6625
6630
6635
6640
6645
6650
6655
6660
6665
6670
6675
6680
6685
6690
6695
6700
6705
6710
6715
6720
6725
6730
6735
6740
6745
6750
6755
6760
6765
6770
6775
6780
6785
6790
6795
6800
6805
6810
6815
6820
6825
6830
6835
6840
6845
6850
6855
6860
6865
6870
6875
6880
6885
6890
6895
6900
6905
6910
6915
6920
6925
6930
6935
6940
6945
6950
6955
6960
6965
6970
6975
6980
6985
6990
6995
7000
7005
7010
7015
7020
7025
7030
7035
7040
7045
7050
7055
7060
7065
7070
7075
7080
7085
7090
7095
7100
7105
7110
7115
7120
7125
7130
7135
7140
7145
7150
7155
7160
7165
7170
7175
7180
7185
7190
7195
7200
7205
7210
7215
7220
7225
7230
7235
7240
7245
7250
7255
7260
7265
7270
7275
7280
7285
7290
7295
7300
7305
7310
7315
7320
7325
7330
7335
7340
7345
7350
7355
7360
7365
7370
7375
7380
7385
7390
7395
7400
7405
7410
7415
7420
7425
7430
7435
7440
7445
7450
7455
7460
7465
7470
7475
7480
7485
7490
7495
7500
7505
7510
7515
7520
7525
7530
7535
7540
7545
7550
7555
7560
7565
7570
7575
7580
7585
7590
7595
7600
7605
7610
7615
7620
7625
7630
7635
7640
7645
7650
7655
7660
7665
7670
7675
7680
7685
7690
7695
7700
7705
7710
7715
7720
7725
7730
7735
7740
7745
7750
7755
7760
7765
7770
7775
7780
7785
7790
7795
7800
7805
7810
7815
7820
7825
7830
7835
7840
7845
7850
7855
7860
7865
7870
7875
7880
7885
7890
7895
7900
7905
7910
7915
7920
7925
7930
7935
7940
7945
7950
7955
7960
7965
7970
7975
7980
7985
7990
7995
8000
8005
8010
8015
8020
8025
8030
8035
8040
8045
8050
8055
8060
8065
8070
8075
8080
8085
8090
8095
8100
8105
8110
8115
8120
8125
8130
8135
8140
8145
8150
8155
8160
8165
8170
8175
8180
8185
8190
8195
8200
8205
8210
8215
8220
8225
8230
8235
8240
8245
8250
8255
8260
8265
8270
8275
8280
8285
8290
8295
8300
8305
8310
8315
8320
8325
8330
8335
8340
8345
8350
8355
8360
8365
8370
8375
8380
8385
8390
8395
8400
8405
8410
8415
8420
8425
8430
8435
8440
8445
8450
8455
8460
8465
8470
8475
8480
8485
8490
8495
8500
8505
8510
8515
8520
8525
8530
8535
8540
8545
8550
8555
8560
8565
8570
8575
8580
8585
8590
8595
8600
8605
8610
8615
8620
8625
8630
8635
8640
8645
8650
8655
8660
8665
8670
8675
8680
8685
8690
8695
8700
8705
8710
8715
8720
8725
8730
8735
8740
8745
8750
8755
8760
8765
8770
8775
8780
8785
8790
8795
8800
8805
8810
8815
8820
8825
8830
8835
8840
8845
8850
8855
8860
8865
8870
8875
8880
8885
8890
8895
8900
8905
8910
8915
8920
8925
8930
8935
8940
8945
8950
8955
8960
8965
8970
8975
8980
8985
8990
8995
9000
9005
9010
9015
9020
9025
9030
9035
9040
9045
9050
9055
9060
9065
9070
9075
9080
9085
9090
9095
9100
9105
9110
9115
9120
9125
9130
9135
9140
9145
9150
9155
9160
9165
9170
9175
9180
9185
9190
9195
9200
9205
9210
9215
9220
9225
9230
9235
9240
9245
9250
9255
9260
9265
9270
9275
9280
9285
9290
9295
9300
9305
9310
9315
9320
9325
9330
9335
9340
9345
9350
9355
9360
9365
9370
9375
9380
9385
9390
9395
9400
9405
9410
9415
9420
9425

" I_{mm} ."

Hybridization intensities for a pair of probes are retrieved at step 954. The background signal intensity is subtracted from each of the hybridization intensities of the pair at step 956. Background subtraction can also be performed on all the raw scan data
5 at the same time.

At step 958, the hybridization intensities of the pair of probes are compared to a difference threshold (D) and a ratio threshold (R). It is determined if the difference between the hybridization intensities of the pair ($I_{pm} - I_{mm}$) is greater than or equal to the difference threshold AND the quotient of the hybridization intensities of the pair (I_{pm} / I_{mm}) is greater than or equal to the ratio threshold. The difference thresholds are
10 typically user defined values that have been determined to produce accurate expression monitoring of a gene or genes. In one embodiment, the difference threshold is 20 and the ratio threshold is 1.2.

If $I_{pm} - I_{mm} \geq D$ and $I_{pm} / I_{mm} \geq R$, the value NPOS is incremented at step 960. In general, NPOS is a value that indicates the number of pairs of probes which have hybridization intensities indicating that the gene is likely expressed. NPOS is utilized in a determination of the expression of the gene.

At step 962, it is determined if $I_{mm} - I_{pm} \geq D$ and $I_{mm} / I_{pm} \geq R$. If these expressions are true, the value NNEG is incremented at step 964. In general,
20 NNEG is a value that indicates the number of pairs of probes which have hybridization intensities indicating that the gene is likely not expressed. NNEG, like NPOS, is utilized in a determination of the expression of the gene.

For each pair that exhibits hybridization intensities either indicating the gene is expressed or not expressed, a log ratio value (LR) and intensity difference value
25 (IDIF) are calculated at step 966. LR is calculated by the log of the quotient of the hybridization intensities of the pair (I_{pm} / I_{mm}). The IDIF is calculated by the difference between the hybridization intensities of the pair ($I_{pm} - I_{mm}$). If there is a next pair of hybridization intensities at step 968, they are retrieved at step 954.

At step 972, a decision matrix is utilized to indicate if the gene is
30 expressed. The decision matrix utilizes the values N, NPOS, NNEG, LR (multiple LRs), and IDIF (multiple IDIFs). The following four assignments are performed:

$$P1 = NPOS / NNEG$$

$$P2 = NPOS / N$$

$$P3 = \text{SUM}(LR) / N$$

$$P4 = \text{SUM}(IDIF)/N$$

These P values are then utilized to determine if the gene is expressed and if the expression level should be displayed. In a preferred embodiment, the expression level of

5 a gene should be displayed if:

$$P1 > 2.2$$

$$P2 > 0.3$$

$$P3 > 0.8$$

$$P4 > 30$$

10 Once all the pairs of probes have been processed and the expression of the gene indicated, an average of the IDIF values for the probes that incremented NPOS or NNEG is calculated at step 975, which is utilized as an expression level. Of course, other values including one of P1 through P4 could be used to indicate expression level.

15 For simplicity, Fig. 5 was described in reference to a single gene or EST. However, the visualization system of the present invention displays expression results for many genes to facilitate discovery of genes of interest or ESTs. Furthermore, the present invention contemplates display of expression levels of a single gene or ESTs as collected from two or more different samples such as tissue samples. The sample sources preferably differ in some characteristic. It will be understood that when the term 20 "sample" is used herein, measurements made on a single "sample" can be based on an aggregation of multiple sample collection events or even multiple organisms.

Fig. 6 shows a screen display illustrating gene expression levels for multiple genes as collected from two tissue samples. A displayed horizontal axis 1002 represents expression level measured in one or more nucleic acid samples taken from the 25 first tissue sample. A displayed vertical axis 1004 represents expression level in one or more nucleic acid samples taken from the second tissue sample. Each of marks 1006 represent a particular gene whose expression level has been measured in both the first and second tissue samples. Each mark 1006 is placed at a distance from vertical axis 1004 corresponding to expression level in the first tissue sample and at a distance from the 30 horizontal axis 1002 corresponding to expression level in the second tissue sample.

The expression levels used for determining the position of marks 1006 are preferably taken from the result of step 975. The position of each of marks 1006 depends on two iterations of the steps of Fig. 5, once for the sample taken from the first tissue

sample and once for the sample taken from the second tissue sample. However, a mark is preferably displayed only if one of the samples meets the threshold criteria at step 972.

In the depicted representative screen display, the first tissue sample is a cancerous tissue sample and the second tissue sample is a normal tissue sample. The 5 individual marks represent the expression levels of selected genes in both cancerous and normal tissue. A first group of marks 1008 represent genes that are neither tumor suppressors nor oncogenes since their expression levels are roughly similar for both normal and cancerous tissue. These marks 1008 fall roughly along a line which is rotated 45 degrees from each of the axes. A second group of marks 1010 represent genes that 10 are likely oncogenes since their expression levels are found to be significantly higher in cancerous tissue than in normal tissue. A third group of marks 1012 represent genes that are likely tumor suppressors since their expression levels are found to be significantly higher in normal tissue than in cancerous tissue. It will be appreciated that expression levels for large numbers of genes can be reviewed at once to discover the oncogenes and tumor suppressors.

Although in the depicted display, the two types of tissue are normal tissue and cancerous tissue, the present invention would aid in the discovery of genes whose expression is associated with any characteristic that varies among tissue samples. For example, once can compare expression results from tissue from individuals who have 20 been exposed to HIV but remain infected to tissue obtained from infected individuals to identify genes conferring resistance to HIV. One can compare expression results between tissue from plants that survive drought to plants that do not. One can compare expression levels among tissue samples at successive stages or severity levels of the same disease, among tissue samples where different ultimate outcomes of the disease (e.g., patient death 25 or remission) are known, among diseased tissue samples that have been subject to different treatment regimes including e.g., chemotherapy, antisense RNA, etc. For cancers, one can compare expression levels between malignant cells and non-malignant cells. Also expression levels can be compared among different organs, between species, and among different stages of development of an organ.

30 It will be appreciated that the present invention also encompasses displays with more than two dimensions. A third visual dimension can be used to illustrate expression level from a third tissue sample. The time dimension can also be used to illustrate successive groups of two or three tissue samples at successive time periods.

The time dimension can be also used to correspond to tissue samples obtained at, e.g., successive stages of a disease.

Other interface methods corresponding to human senses other than sight can also be incorporated within the presentation system of the present invention. The 5 senses may correspond to additional dimensions. For example, marks can be displayed in succession accompanied by a sound having characteristics corresponding to expression level in another tissue sample.

The user can employ a cursor 1014 to identify a particular mark as being of interest. Cursor 1014 can be moved to a particular mark by use of, e.g., mouse 11. 10 Once cursor 1014 is over a mark of interest, the mark can be selected by, e.g., depression of one of mouse buttons 13. Selection of a particular mark can be facilitated by use of a zoom display feature (not shown). Once a particular mark is selected, further information is displayed about the gene represented by the mark. A special mouse can transmit a tactile sensation back to the user corresponding to expression level in a tissue 15 sample as the user passes the mouse over a corresponding mark.

It will be appreciated that the display of Fig. 6 is not limited to expression information. The two dimensions of Fig. 6 may correspond to indicators of the presence of various polymers other than nucleic acids in two different samples. For example, each mark may correspond to a different polymer, polypeptide, or other compound. The 20 distance of the mark from each axis would correspond to a measure of presence of the particular polymer in the sample corresponding to the axis. One possible measure is produced by fluorescently tagging polymer samples such as protein samples and exposing a probe array such as a peptide probe array to the protein samples. The fluorescent intensity of the probes will then correspond to the bonding affinity of the sample to the 25 probes. The intensity measurement or a measurement derived from the intensity measurement may then be used to position the marks of Fig. 6.

Fig. 7A shows a screen display giving information about a particular gene selected from the display of Fig. 6. A cluster number 702, a GenBank accession number 704, and a verbal description 706 for the selected gene are displayed. The user can also 30 select a number of marks 1006 by circling them with cursor 1014. Then a list of information as shown in Fig. 7A is displayed for all the genes corresponding to the selected marks.

By selecting GenBank accession number 704 with another cursor (not

shown), the user can direct retrieval of the GenBank information for the selected gene. If the GenBank information is not available locally, the retrieval process can include formulating a query and transmitting the query to a GenBank web site. Once the GenBank information is retrieved, it can also be displayed. Fig. 7B depicts the GenBank

5 information for the gene identified in Fig. 7A.

In the foregoing specification, the invention has been described with reference to specific exemplary embodiments thereof. It will, however, be evident that various modifications and changes may be made thereunto without departing from the broader spirit and scope of the invention as set forth in the appended claims and their full scope of equivalents.

10
10028748-162403